On a Class of Shrinkage Priors for Covariance Matrix Estimation

Hao Wang

Department of Statistics, University of South Carolina,
Columbia, SC 29208, U.S.A.
haowang@sc.edu

Natesh S. Pillai

Department of Statistics, Harvard University,
Cambridge, MA 02138, U.S.A.
pillai@fas.harvard.edu

This version: October 7, 2011

Abstract

We propose a flexible class of models based on scale mixture of uniform distributions to construct shrinkage priors for covariance matrix estimation. This new class of priors enjoys a number of advantages over the traditional scale mixture of normal priors, including its simplicity and flexibility in characterizing the prior density. We also exhibit a simple, easy to implement Gibbs sampler for posterior simulation which leads to efficient estimation in high dimensional problems. We first discuss the theory and computational details of this new approach and then extend the basic model to a new class of multivariate conditional autoregressive models for analyzing multivariate areal data. The proposed spatial model flexibly characterizes both the spatial and the outcome correlation structures at an appealing computational cost. Examples consisting of both synthetic and real-world data show the utility of this new framework in terms of robust estimation as well as improved predictive performance.

Key words: Areal data; Covariance matrix; Data augmentation Gibbs sampler; Multivariate conditional autoregressive model; Scale mixture of uniform; Shrinkage; Sparsity.

1 Introduction

Estimation of the covariance matrix Σ of a multivariate random vector y is ubiquitous in modern statistics and is particularly challenging when the dimension of the covariance matrix, p, is comparable or even larger than the sample size n. For efficient inference, it is thus paramount to take advantage of parsimonious structure often inherent in these high dimensional problems. Many Bayesian approaches have been proposed for covariance matrix estimation by placing shrinkage priors on various parameterizations of the covariance matrix Σ. Yang & Berger (1994) proposed reference priors for Σ based on the spectral decomposition of Σ. Barnard et al. (2000) and Liechty et al. (2004)

considered shrinkage priors in terms of the correlation matrix and standard deviations. Daniels & Kass (1999, 2001) proposed flexible hierarchical priors based on a number of parameterizations of Σ. All of these methods use non-conjugate priors and typically rely on Markov chain algorithms which explore the state space locally such as Metropolis-Hastings methods or asymptotic approximations for posterior simulation and modeling fitting and are restricted to low-dimensional problems.

A large class of sparsity modeling of the covariance matrix involves the identification of zeros of the inverse $\Omega = \Sigma^{-1}$. This corresponds to the Gaussian graphical models in which zeros in the inverse covariance matrix uniquely determine an undirected graph that represents the strict conditional independencies. The Gaussian graphical model approach for covariance matrix estimation is attractive and has gained substantive attention owing to the fact that its implied conditional dependence structure provides a natural platform for modeling dependence of random quantities in areas such as biology, finance, environmental health and social sciences. The standard Bayesian approach to inference in Gaussian graphical models is the conjugate G-Wishart prior (Roverato, 2002; Atay-Kayis & Massam, 2005), which places positive probability mass at zero on zero elements of Ω. A zero constrained random matrix Ω has the G-Wishart distribution $\mathrm{W}_G(b, D)$ if its density is

$$p(\Omega \mid G) \;=\; C_G(b,D)^{-1}|\Omega|^{(b-2)/2}\exp\{-\frac{1}{2}\mathrm{tr}(D\Omega)\}\,1_{\{\Omega\in M^+(G)\}}, \tag{1}$$

where $b > 2$ is the degree of freedom parameter, D is a symmetric positive definite matrix, $C_G(b, D)$ is the normalizing constant, $M^+(G)$ is the cone of symmetric positive definite matrices with entries corresponding to the missing edges of G constrained to be equal to zero, and $1_{\{\cdot\}}$ is the indicator function. Although G-Wishart prior has been quite successfully used in many applications, it has a few important limitations. First, the G-Wishart prior is sometimes not very flexible because of its restrictive form. For example, the parameters for the degrees of freedom are the same for all the elements of Ω. Second, unrestricted graphical model determination and covariance matrix estimation is computationally challenging. Recent advances for unrestricted graphical models (Jones et al., 2005; Wang & Carvalho, 2010; Mitsakakis et al., 2010; Dobra et al., 2011) all rely on the theoretical framework of Atay-Kayis & Massam (2005) for sparse matrix completion which is very computationally intensive. Indeed, for non-decomposable graphical models, we do not have a closed form expression for the normalizing constant $C_G(b, D)$ and thus have to resort to tedious and often unstable Monte Carlo integration to estimate it for both graphical model determination and covariance matrix estimation.

An alternative method for Bayesian graphical model determination and estimation is proposed by Wong et al. (2003). They placed point mass priors at zero on zero elements of the partial correlation matrix and constant priors for the non-zero elements. Their methodology applies to both decomposable and non-decomposable models and is fitted by a reversible jump Metropolis-Hastings algorithm. However, it is unclear how to incorporate prior information about individual entries of Σ in their framework as the mathematical convenience of constant priors is essential for their algorithm.

Absolutely continuous priors, or equivalently, penalty functions, can also induce shrinkage to zero of subsets of elements of Ω and represent an important and flexible alternative to the point mass priors. In the classical formulation, there is a rich literature on methods for developing shrinkage estimators via different penalty functions including the graphical lasso models (Yuan & Lin, 2007; Friedman et al., 2008; Rothman et al., 2008) and the graphical adaptive lasso models (Fan et al., 2009) among many others. The recent literature on Bayesian methods has focused on the posterior mode estimation, with little attention on the key problem of efficient inference on covariance matrix based on full posterior computation, with the only exception of Wang (2011) which gave a fully Bayesian treatment of the graphical lasso models. One likely reason is the difficulty in efficiently generating posterior samples of covariance matrices under shrinkage priors. A fully Bayesian inference is quite desirable because it not only produces valid standard errors and Bayes estimators based on decision-theoretic framework but, perhaps more importantly, can be applied in multiple classes of multivariate models that involve key components of unknown covariance matrices such as the multivariate conditional autoregressive models developed in Section 5.

This paper proposes a class of priors and the implied Bayesian hierarchical modeling and computation for shrinkage estimation of covariance matrices. A key but well known observation is that any symmetric, unimodal density may be written as a scale mixture of uniform distributions. Our main strategy is to use this mixture representations to construct shrinkage priors compared to the traditional methods for constructing shrinkage priors using the scale mixture of normal distributions. As mentioned above, the scale mixture of uniform distribution is not new to Bayesian inference. Early usage of this representation includes Bayesian robust and sensitive analysis (Berger, 1985; Berger & Berliner, 1986) and robust regressions with heavy-tailed errors Walker et al. (1997).

However, our motivations are different; we seek an approach for constructing tractable shrinkage priors that are both flexible and computationally efficient. We argue that the class of scale mixture of uniform priors provide an appealing framework for modeling a wide class of shrinkage estimation problems and also has the potential to be extended to a large class of high dimensional problems involving multivariate dependencies. We also highlight that a salient feature of our approach is its computational simplicity. We construct a simple, easy to implement Gibbs sampler based on data augmentation for obtaining posterior draws for a large class of shrinkage priors. To the best of our knowledge, none of the existing Bayesian algorithms for sparse permutation invariant covariance estimation can be carried out solely based on a Gibbs sampler and they have to rely on Metropolis-Hastings methods. Since Gibbs samplers involve global proposal moves as compared to the local proposals of Metropolis-Hastings methods, in high dimensions this makes a difference in both the efficiency of the sampler and the running time of the algorithm. Through simulation experiments, we illustrate the robust performance of the scale mixture of uniform priors for covariance matrix, as well as highlighting the strength and weakness of this approach compared to those based on point mass priors. Through an extension to a class of multivariate conditional autoregressive models, we further illustrate that the framework of scale mixture of uniforms naturally allows and encourages the integration of data and expert knowledge in model fitting and assessment, and consequently improves the prediction.

The rest of the paper is organized as follows. In Section 2 we outline our framework for constructing shrinkage priors for covariance matrices using the scale mixture of uniforms. In Section 3 we construct a Gibbs sampler based on a data augmentation scheme for sampling from the posterior distribution. In Section 4 we conduct a simulation study and compare and contrast our models with existing methods. In Section 5 we extend our model to build shrinkage priors on multivariate conditional autoregressive models. In Section 6 we briefly discuss the application of our methods for shrinkage estimation for regression models.

2 Shrinkage priors for precision matrices

2.1 Precision matrix modeling

Let $y = (y^{(1)}, y^{(2)}, \ldots, y^{(p)})^{\mathrm{T}}$ be a p-dimensional random vector having a multivariate normal distribution $\mathrm{N}(0, \Sigma)$ with mean zero and covariance matrix Σ. Let $\Omega = (\omega_{ij})_{p \times p} = \Sigma^{-1}$ denote the precision matrix, i.e., the inverse of the covariance matrix Σ. Given a set of independent random samples $Y = (y_1, \ldots, y_n)_{p \times n}$ of y, we wish to estimate the matrix Ω.

We consider the following prior distribution for the precision matrix:

$$p(\Omega \mid \tau) \quad \propto \quad \prod_{i \leq j} g_{ij}\left(\frac{\omega_{ij} - m_{ij}}{\tau_{ij}}\right) 1_{\Omega \in M^+}, \tag{2}$$

where $g_{ij}(\cdot)$ is a continuous, unimodal and symmetric probability density function with mode zero on $\mathbb{R}$, M^+ is the space of real valued symmetric, positive definite $p \times p$ matrices, $\tau_{ij} > 0$ is a scale parameter controlling the strength of the shrinkage and 1_A denotes the indicator function of the set A. Our primary motivation for constructing prior distributions of the form (2) is that, often in real applications the amount of prior information the modeler can vary across individual elements of Ω. For instance, one might incorporate the information that the variance of certain entries of Ω are close to 0, or constrain some entries to be exactly 0. In this setting, shrinking different elements of Ω at a different rate clearly provides a flexible framework for conducting Bayesian inference. In addition to obtaining a flexible class of prior distributions, by using a mixture representation for the density g_{ij}, we can construct a simple, efficient and easy to implement Gibbs sampler to draw from the posterior distribution of the precision matrix.

2.2 Scale mixture of uniform distributions

Our main tool is the following theorem which says that all unimodal, symmetric densities may be expressed as scale mixture of uniform distributions.

Theorem 1. *Walker et al. (1997); Feller (1971) Suppose that θ is a real-valued random quantity with a continuous, unimodal and symmetric distribution with mode zero having density $\pi(\theta)\,(-\infty < \theta < \infty)$. Suppose $\pi'(\theta)$ exists for all θ. Then $\pi(\theta)$ has the form:*

$$\pi(\theta) \quad = \quad \int_0^\infty \frac{1}{2t}\, 1_{\{|\theta|<t\}}\, h(t)\, \mathrm{d}t, \tag{3}$$

where $h(t) \propto -2t \times \pi'(t)$ is some density function on $[0,\infty)$. Therefore we may write

$$\pi(\theta \mid t) \sim \mathrm{U}(-t,t), \quad h(t) \propto -2t \times \pi'(t).$$

The generality and simplicity of Theorem 1 allow us to characterize various shrinkage priors by using the special structure of (3). Indeed, as noted in Walker et al. (1997), a Gaussian random variable $x \sim \mathrm{N}(\mu,\sigma^2)$ can be expressed as $x \mid v \sim \mathrm{U}(\mu - \sigma\sqrt{v}, \mu + \sigma\sqrt{v}), v \sim \mathrm{G\small{A}}(3/2, 1/2)$, which shows that, all of the distributions which may be written as a scale mixture of Gaussian distributions can indeed be expressed as a scale mixture of uniform distributions as well. Let us discuss a few more examples of popular shrinkage priors where Theorem 1 is applicable.

A popular class of distributions for constructing shrinkage priors is the exponential power family given by $\pi(\theta) \propto \exp(-|\theta|^q/\tau^q)$, where the exponent $q > 0$ controls the decay at the tails. The mixing density function $h(t)$ given in (3) can be thought of as the "scale" parameter. In this case we have $h(t) \propto t^q \exp(-t^q/\tau^q)$, which corresponds to the generalized gamma distribution. Two important special cases are the Gaussian distribution ($q = 2$), and the double-exponential distribution ($q = 1$), which have been studied extensively in the context of the Bayesian lasso regression (Park & Casella, 2008; Hans, 2009) and the Bayesian graphical lasso (Wang, 2011). For general $q > 0$, one may write the exponential power distribution as a scale mixture of Gaussian distributions (Andrews & Mallows, 1974; West, 1987). However, a fully Bayesian, computationally efficient analysis is not available based on Gaussian mixtures, especially in the context of covariance estimation and graphical models. A few approximate methods exist for doing inference using the exponential power prior distribution such as the variational method proposed by Armagan (2009). Our use of uniform mixture representation has the advantage of posterior simulation via an efficient Gibbs sampler for any $q > 0$ as is shown in Section 2.3 and further exemplified in Sections 4 and 6.

Another natural candidate for shrinkage priors is the Student-t distribution given by $\pi(\theta) \propto (1+\theta^2/\tau^2)^{-(\nu+1)/2}$, for which it is easy to show that $h(t) \propto t^2(1+t^2/\tau^2)^{-(\nu+3)/2}$. Hence, t^2/τ^2 is an inverted beta distribution $\mathrm{IB}(3/2, \nu/2)$. Recall that the inverted beta distribution $\mathrm{IB}(a,b)$ has the density given by $p(x) \propto x^{a-1}(1+x)^{-a-b}1_{x>0}$.

The generalized double Pareto distribution is given by $\pi(\theta) \propto (1 + |\theta|/\tau)^{-(1+\alpha)}$, which corresponds to $h(t) \propto t(1+t/\tau)^{-(2+\alpha)}$; i.e., the scale t/τ follows an inverted beta distribution $\mathrm{IB}(2, \alpha)$. Armagan et al. (2011) investigated the properties of this class of shrinkage priors.

The above discussed class of shrinkage priors are well known and documented. In the following we give a new distribution which we call the "logarithmic" shrinkage prior which seems to be new in the context of shrinkage priors. Consider the density given by

$$\pi(\theta) \;\; \propto \;\; \log(1 + \tau^2/\theta^2) \; . \tag{4}$$

It is easy to show that the corresponding mixing distribution has the half-Cauchy density,

$$h(t) \propto (1 + t^2/\tau^2)^{-1} 1_{\{t>0\}} \; .$$

This prior has two desirable properties for shrinkage estimation: an infinite spike at zero and heavy tails. These are precisely the desirable characteristics of a shrinkage prior distribution as argued convincingly for the "horseshoe" prior in Carvalho et al. (2010). The horseshoe prior is constructed by scale mixture of normals, namely, $\theta \sim \mathrm{N}(0, \sigma^2), \sigma \sim \mathrm{C}^+(0, 1)$, where $\mathrm{C}^+(0, 1)$ is a standard half-Cauchy distribution on the positive reals with scale 1. The horseshoe prior does not have a closed form density but satisfies the following:

$$\frac{K}{2} \log(1 + 4/\theta^2) < \pi(\theta) < K \log(1 + 2/\theta^2),$$

for a constant $K > 0$. Clearly, our new prior (4) has identical behavior at the original and the tails as that of the horseshoe prior distribution with the added advantage of having an explicit density function unlike the horseshoe prior.

2.3 Posterior sampling

Let y denote the observed data. The scale mixture of uniform representation provides a simple way of sampling from the posterior distribution $p(\theta \mid y) \propto f(y \mid \theta)\pi(\theta)$, where $f(y \mid \theta)$ is the likelihood function and $\pi(\theta)$ is the shrinkage prior density. The representation (3) leads to the following full conditional distributions of θ and t (conditional on y) given by

$$p(\theta \mid y, t) \propto f(y \mid \theta) \, 1_{|\theta|<t}, \quad p(t \mid y, \theta) \propto -\pi'(t) \, 1_{|\theta|<t} \; . \tag{5}$$

Thus the data augmented Gibbs sampler for obtaining posterior draws from $p(\theta, t \mid y)$ involves iteratively simulating from the above two conditional distributions. Simulation of the former involves sampling from a truncated distribution, which is often achieved by breaking it down further into several Gibbs steps, while sampling the latter is achieved by the following theorem.

Table 1: Density of θ and t for some common shrinkage prior distributions, along with the conditional posterior inverse cumulative probability function for sampling t. Densities are given up to normalizing constants.

Density name	Density for θ	Density for t	Inverse CDF: $F^{-1}(u \mid \theta)$
Exponential power	$\exp(-\lvert\theta\rvert^q/\tau^q)$	$t^q \exp(-t^q/\tau^q)$	$\{-\tau^q(\log u) + \lvert\theta\rvert^q\}^{1/q}$
Student-t	$(1+\theta^2/\tau^2)^{-(\nu+1)/2}$	$t^2(\nu + t^2/\tau^2)^{-(\nu+3)/2}$	$\{u^{-2/(\nu+1)}(\tau^2+\theta^2) - \tau^2\}^{1/2}$
Generalized double Pareto	$(1+\lvert\theta\rvert/\tau)^{-(1+\alpha)}$	$t(1+t/\tau)^{-(2+\alpha)}$	$u^{-1/(1+\alpha)}(\lvert\theta\rvert+\tau) - \tau$
Logarithmic	$\log(1+\tau^2/\theta^2)$	$(1+t^2/\tau^2)^{-1}$	$\tau\{(1+\tau^2/\theta^2)^u - 1\}^{-1/2}$

CDF, cumulative distribution function.

Theorem 2. *Suppose the shrinkage prior density $\pi(\theta)$ can be represented by a scale mixture of uniform as in equation (3). Then the (posterior) conditional probability density function of the latent scale parameter t is given by*

$$p(t \mid y, \theta) \propto -\pi'(t)\, 1_{t>|\theta|},$$

and the corresponding (conditional) cumulative distribution function is

$$u = F(t \mid y, \theta) = pr(T < t \mid y, \theta) \quad = \quad \frac{\pi(|\theta|) - \pi(t)}{\pi(|\theta|)} \quad |\theta| < t \,. \tag{6}$$

The advantage of the above theorem is that it gives an explicit expression of the conditional cumulative distribution function in terms of the prior density $\pi(\cdot)$. This provides a simple way to sample from $p(t \mid y, \theta)$ using the inverse cumulative distribution function method whenever $\pi(\cdot)$ can be easily inverted. Table 1 summarizes the density functions of $\pi(\theta)$ and $h(t)$, and the inverse conditional cumulative distribution function $F^{-1}(u \mid y, \theta)$ for several shrinkage priors introduced in Section 2.2. We note that the scale mixture of uniform distributions are already used for doing inference for regression models using the Gibbs sampler outlined above, for instance see Qin et al. (1998).

3 Posterior computation for precision matrices

3.1 Gibbs sampling on given global shrinkage parameter τ

Recall that given a set of independent random samples $Y = (y_1, \dots, y_n)_{p\times n}$ from a multivariate normal distribution $\mathrm{N}(0, \Omega^{-1})$, we wish to estimate the matrix Ω using the prior distribution given by (2). Let $T = \{t_{ij}\}_{i\geq j}$ be the vector of latent scale parameters. For simplicity we first consider a simple case where $g_{ij}(\cdot) = g(\cdot), m_{ij} = 0$ and $\tau_{ij} = \tau$ in this section, and then discuss the strategies for choosing τ in Section 3.2. However our algorithms can be easily extended to the general case of unequal shrinkage parameters τ_{ij}. Theorem 1 suggests that the prior (2) can be represented as follows:

$$p(\Omega \mid \tau) = \int_T p(\Omega, T \mid \tau)\mathrm{d}T \propto \int_T \prod_{i\geq j} \big[\frac{1}{2t_{ij}} 1_{\{|\omega_{ij}| < \tau t_{ij}\}} h(t_{ij})\big]\mathrm{d}T,$$

where $p(\Omega, T \mid \tau) \propto \prod_{i\geq j} \left[1/(2t_{ij})\, 1_{\{|\omega_{ij}|<\tau t_{ij}\}} h(t_{ij})\right]$ is the joint prior and $h(t_{ij}) \propto -t_{ij} g'(t_{ij})$. The joint posterior distribution of (Ω, T) is then:

$$p(\Omega, T \mid Y, \tau) \;\; \propto \;\; |\Omega|^{n/2} \exp\{-\frac{1}{2}\text{tr}(S\Omega)\} \prod_{i\geq j} \left[-1_{\{|\omega_{ij}|<\tau t_{ij}\}}\, g'(t_{ij})\right], \tag{7}$$

where $S = YY^{\mathrm{T}}$.

The most direct approach for sampling from (7) is to update each ω_{ij} one at a time given the data, T, and all of the entries in Ω except for ω_{ij} in a way similar to those proposed in Wong et al. (2003). However, this direct approach requires a separate Cholesky factorization for updating each ω_{ij} to find its allowable range and conditional distribution. It also relies on the Metropolis-Hastings step to correct the sample. We describe an efficient Gibbs sampler for sampling (Ω, T) from (7) that involves one step for sampling Ω and the other step for sampling T.

Given T, the first step of our Gibbs sampler systematically scans the set of 2×2 sub-matrices $\{\Omega_{e,e} : e = (i,j), 1 \leq j < i \leq p\}$ to generate Ω. For any $e = (i,j)$, let $V = \{1, \ldots, p\}$ be the set of vertices and note that

$$|\Omega| = |A||\Omega_{V\setminus e, V\setminus e}|,$$

where A, the Schur component of $\Omega_{V\setminus e, V\setminus e}$, is defined by $A = \Omega_{e,e} - B$ with $B = \Omega_{e,V\setminus e}(\Omega_{V\setminus e,V\setminus e})^{-1}\Omega_{V\setminus e,e}$. The full conditional density of $\Omega_{e,e}$ from (7) is given by

$$p(\Omega_{e,e} \mid -) \propto |A|^{n/2} \exp\{-\frac{1}{2}S_{e,e}A\}\, 1_{\{\Omega_{e,e}\in\mathcal{T}\}},$$

where $\mathcal{T} = \{|\omega_{ij}| < \tau t_{ij}\} \cap \{|\omega_{ii}| < \tau t_{ii}\} \cap \{|\omega_{jj}| < \tau t_{jj}\}$. Thus, A is a truncated Wishart variate. To sample A, we write

$$A = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & l_{21} \\ 0 & 1 \end{pmatrix}, \quad S_{e,e} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix},$$

with $d_1 > 0$ and $d_2 > 0$. The joint distribution of (l_{12}, d_1, d_2) is then given by:

$$p(d_1, d_2, l_{21} \mid -) \propto d_1^{n/2+1} d_2^{n/2} \exp[-\frac{1}{2}\text{tr}\{s_{11}d_1 + s_{22}(l_{21}^2 d_1 + d_2) + 2s_{21}d_1 l_{21}\}]\, 1_{\Omega_{e,e}\in\mathcal{T}},$$

which implies that the univariate conditional distribution for the parameters d_1 and d_2 is a truncated gamma distribution, and a truncated normal distribution for l_{21}. Details of the parameters of the truncated region and strategies for sampling are given in the Appendix. Given Ω, the second step of our Gibbs sampler generates T in block using the inverse cumulative distribution function methods described in equation (6). These two steps complete a Gibbs sampler for model fitting under a broad class of shrinkage priors for Ω.

One attractive feature of the above sampler is that it is also suitable for sampling $\Omega \in M^+(G)$, that is, Ω is constrained by an undirected graph $G = (V, E)$ where V is the set of vertices and E is a set of edges and $\omega_{ij} = 0$ if and only if $(i,j) \notin E$. The ability to sample $\Omega \in M^+(G)$ is useful when substantive prior information indicates a certain subset of elements in Ω are indeed zero. Section 5 provides such an example that involves a class of multivariate spatial models. To sample $\Omega \in M^+(G)$, the only modification that is required is to replace the set of all 2×2 sub-matrices $\{\Omega_{e,e} : e = (i,j), 1 \leq j < i \leq p\}$ with the set $\{\Omega_{e,e} : e \in E\} \cup \{\Omega_v : v \in V_I\}$ where V_I is the set of isolated nodes in G.

3.2 Choosing the shrinkage parameters

We start with the scenario when $\tau_{ij} = \tau$ and $m_{ij} = 0$ for all $i \geq j$. In this case we have

$$p(\Omega \mid \tau) = C_\tau^{-1} \prod_{i \geq j} g(\frac{\omega_{ij}}{\tau}),$$

where C_τ is a normalizing term involving τ. This normalizing constant is a necessary quantity for choosing hyper parameters for τ. Since $p(\Omega \mid \tau)$ is a scale family, applying the substitution $\tilde{\Omega} = \Omega/\tau$ yields,

$$C_\tau = \int_{\Omega \in M^+} \prod_{i \geq j} g(\frac{\omega_{ij}}{\tau}) \mathrm{d}\Omega = \tau^{\frac{p(p+1)}{2}} \int_{\tilde{\Omega} \in M^+} g(\tilde{\omega}_{ij}) \mathrm{d}\tilde{\Omega}, \tag{8}$$

where the integral on the right hand side of the above equation does not involve τ because $\{\tilde{\Omega} : \tilde{\Omega} \in M^+\} = \{\Omega : \Omega \in M^+\}$. Hence, under a hyperprior $p(\tau)$, the conditional posterior distribution of τ is

$$p(\tau \mid Y, \Omega) \propto \tau^{-p(p+1)/2} \prod_{i \geq j} g(\frac{\omega_{ij}}{\tau}) p(\tau) \; . \tag{9}$$

Now the sampling scheme in Section 3.1 can be extended to include a component to sample τ at each iteration.

Now suppose $m_{ij} = 0$ and instead of having a single global shrinkage parameter, we wish to control the rate at which the individuals elements of Ω are shrunk towards 0 separately. A natural shrinkage prior for this problem is

$$p(\Omega \mid \tau) = C_\tau^{-1} \prod_{i \geq j} g_{ij}(\frac{\omega_{ij}}{\tau})$$

where g_{ij} may all be different. The idea is that by choosing a different density g_{ij} for each edge, we can incorporate the prior information for the rate at which different entries of Ω are shrunk towards 0. For a hyper prior $p(\tau)$, using an identical calculation as in (8) and (9) we deduce that the conditional posterior of τ is then given by

$$p(\tau \mid Y, \Omega) \propto \tau^{-p(p+1)/2} \prod_{i \geq j} g_{ij}(\frac{\omega_{ij}}{\tau}) p(\tau) \; . \tag{10}$$

Notice that the Gibbs sampler presented in Section 3.1 applies to this case as well; we just need to use the cumulative distribution function for the density g_{ij} for sampling from the conditional distribution of t_{ij}. Alternatively, one can also fix a density g and write $p(\Omega \mid \tau) = C_\tau^{-1} \prod_{i \geq j} g(\frac{\omega_{ij}}{v_{ij}\tau})$ for fixed positive constants v_{ij} and then make inference about the common τ.

We conclude this section with the remark that our approach can be adapted for hierarchical models. For example, in Section 5 we consider a shrinkage prior that shrinks Ω towards a given matrix $M = (m_{ij})$ under the constraint that $\Omega \in M^+(G)$ for a given graph G:

$$p(\Omega) = C_{\tau,M}^{-1} \prod_{(i,j)\in E} g(\frac{\omega_{ij} - m_{ij}}{\tau}) 1_{\Omega \in M^+(G)},$$

where E denotes the set of edges of the graph G and normalizing constant $C_{\tau,M} = \int_{\Omega \in M^+(G)} \prod_{(i,j)\in E} g(\frac{\omega_{ij}-m_{ij}}{\tau}) \mathrm{d}\Omega$ is the normalizing constant. In this case $C_{\tau,M}$ is analytically intractable as a function of τ. In the example of Section 5, we fixed τ at a value that represents prior knowledge of the distribution of Ω to avoid modeling τ. In some applications, it may be desirable to add another level of hierarchy for modeling τ so that we can estimate it from data. Several approaches have been proposed for dealing with the intractable normalizing constant, see Liechty et al. (2004), Liechty et al. (2009) and the references therein for one such approach.

4 Simulation experiments

To assess the utility of the scale mixture of uniform priors, we compared a range of priors in this family against three alternatives: the frequentist graphical lasso method of Friedman et al. (2008), the Bayesian G-Wishart prior and the method of Wong et al. (2003). The latter two place positive prior mass on zeros. We considered four covariance matrices from Rothman et al. (2008):

***Model* 1.** *An* AR*(1) model with* $\sigma_{ij} = 0{\cdot}7^{|i-j|}$.

***Model* 2.** *An* AR*(4) model with* $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0{\cdot}2$, $\omega_{i,i-2} = \omega_{i-2,i} = \omega_{i,i-3} = \omega_{i-3,i} = 0{\cdot}2$, $\omega_{i,i-4} = \omega_{i-4,i} = 0{\cdot}1$.

***Model* 3.** *A sparse model with* $\Omega = B + \delta I_p$ *where each off-diagonal entry in* B *is generated independently and assigned the value 0·5 with probability* $\alpha = 0{\cdot}1$ *and 0 otherwise. The diagonal elements* B_{ii} *are set to be 0, and* δ *is chosen so that the condition number of* Ω *is* p. *Here the condition number is defined as* $\max(\lambda)/\min(\lambda)$ *where* $\max(\lambda), \min(\lambda)$ *respectively denote the maximum and minimum eigenvalues of the matrix* Ω.

***Model* 4.** *A dense model with the same* Ω *as in model 3 except for* $\alpha = 0{\cdot}5$.

For each of the above four models, we generated samples of size $n = 30, 100$ and dimension $p = 30$, yielding the proportion of non-zero elements to be $6\%, 25\%, 10\%, 50\%$, respectively. We compute the risk under two standard loss functions, Stein's loss function, $L_1(\hat{\Sigma}, \Sigma) = \mathrm{tr}(\hat{\Sigma}\Sigma^{-1}) - \log(\hat{\Sigma}\Sigma^{-1}) - p$, and the squared-error loss function $L_2(\hat{\Sigma}, \Sigma) = \mathrm{tr}(\hat{\Sigma} - \Sigma)^2$. The corresponding Bayes estimators are $\hat{\Sigma}_{L_1} = \{\mathrm{E}(\Omega \mid \mathrm{Y})\}^{-1}$ and $\hat{\Sigma}_{L_2} = \mathrm{E}(\Sigma \mid \mathrm{Y})$, respectively. We used the posterior sample mean using the Gibbs sampler for estimating the risk for the Bayesian methods and the maximum likelihood estimate for the graphical lasso method.

When fitting graphical lasso models, we used the 10-fold cross-validation to choose the shrinkage parameter. When fitting the G-Wishart priors, we followed the conventional prior specification $\Omega \sim \mathrm{W}_G(3, I_p)$ and used the reversible jump algorithm of Dobra et al. (2011) for model fitting. For both the G-Wishart priors and the methods of Wong et al. (2003), we used the default graphical model space prior (Carvalho & Scott, 2009)

$$p(G) = \{(1+m)\binom{m}{|G|}\}^{-1},$$

where $m = p(p-1)/2$ and $|G|$ is the total number of edges in graph G. For the scale mixtures of uniforms, we considered the exponential power prior $p(\Omega \mid \tau) \propto \exp\{-\sum_{i\leq j} |\omega_{ij}|^q/\tau^q\}$ with $q \in \{0{\cdot}2, 1\}$, the generalized double-Pareto prior $p(\Omega \mid \tau) \propto \prod_{i\leq j}(1+|\omega_{ij}|/\tau)^{-1-\alpha}$ and the new logarithmic prior $p(\Omega \mid \tau) \propto \prod_{i\leq j} \log(1+\tau^2/\omega_{ij}^2)$. For the choice of the global shrinkage parameters, we assumed the conjugate distribution $\tau^{-q} \sim \mathrm{Ga}(1,\, 0{\cdot}1)$ for the exponential power prior; $\alpha = 1, 1/(1+\tau) \sim \mathrm{U}(0,1)$ for the generalized double Pareto prior as suggested by Armagan et al. (2011); and $\tau \sim \mathrm{C}^+(0,1)$ for the logarithmic prior as was done for the horseshoe prior in Carvalho et al. (2010).

Twenty datasets were generated for each case. The Bayesian procedures used 15000 iterations with the first 5000 as burn-ins. In all cases, the convergence was rapid and the mixing was good; the autocorrelation of each elements in Ω died out typically after 10 lags. As for the computational cost, the scale mixture of uniforms and the method of Wong et al. (2003) were significantly faster than the G-Wishart method. For example, for model 4, the G-Wishart took about 11 hours for one dataset under Matlab implementation on a six core 3·3 Ghz computer running CentOS 5·0 unix ; while the scale mixture of uniforms and the method of Wong et al. (2003) took only about 20 and 6 minutes respectively. The graphical lasso method is just used as a benchmark for calibrating the Bayesian procedures. For each dataset, all Bayesian methods were compared to the graphical lasso method by computing the relative loss; for example, for the L_1 loss, we computed the relative loss as $L_1(\hat{\Sigma}, \Sigma) - L_1(\hat{\Sigma}_{\mathrm{GLASSO}}, \Sigma)$, where $\hat{\Sigma}$ is any Bayes estimator of Σ and Σ_{GLASSO} is the graphical lasso estimator. Thus, a negative value indicates that the method performs better relative to the graphical lasso procedure and a smaller relative loss indicates a better relative performance of the method.

Table 2 reports the simulation results. The two approaches based on point mass priors outperform the continuous shrinkage methods in sparser models such as model 1, however, they are outperformed in less sparse configurations such as model 2 and 4. One possible explanation is that the point mass priors tend to favor sparse models because it encourages sparsity through a positive prior mass at zero. Finally, the exponential power with q =0·2, the generalized double Pareto and the logarithmic

Table 2: Summary of the relative L_1 and L_2 losses for different models and different methods. Medians are reported while standard errors are in parentheses.

		Model 1		Model 2		Model 3		Model 4	
		L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
n=30	W_G	-4·4 (1·3)	-5·9 (1·4)	-0·3 (0·7)	-12·7 (4·6)	-0·9 (0·7)	1·4 (2·5)	-2·3 (1·9)	-0·0 (0·9)
	WCK	-4·4 (1·0)	-5·1 (2·3)	-0·7 (0·6)	-11·3 (3·8)	-1·2 (0·6)	1·6 (1·5)	-2·2 (1·0)	0·3 (0·5)
	$\mathrm{EP}_{q=1}$	-2·1 (1·1)	2·1 (1·0)	-1·0 (0·8)	-14·0 (4·7)	-1·6 (0·7)	-1·0 (2·2)	-4·2 (1·2)	-1·1 (0·5)
	$\mathrm{EP}_{q=0·2}$	-3·8 (1·1)	-2·9 (2·1)	-0·9 (0·8)	-13·7 (4·9)	-1·4 (0·7)	-0·5 (2·5)	-3·1 (1·1)	-0·5 (1·3)
	GDP	-3·8 (1·1)	-3·2 (2·2)	-1·3 (0·7)	-13·2 (4·3)	-1·4 (0·7)	-0·4 (2·3)	-2·5 (1·7)	-0·4 (0·9)
	Log	-3·7 (1·1)	-2·3 (1·4)	-0·6 (0·8)	-13·3 (4·9)	-1·3 (0·6)	-0·2 (2·5)	-3·2 (1·1)	-0·8 (0·9)
n=100	W_G	-1·7 (0·2)	-3·9 (0·7)	-0·3 (0·4)	-0·4 (1·5)	-0·8 (0·3)	-1·5 (1·5)	0·4 (0·3)	0·7 (0·3)
	WCK	-1·3 (0·2)	-2·7 (0·6)	-0·7 (0·3)	-0·8 (1·1)	-0·5 (0·2)	0·3 (1·4)	0·2 (0·3)	0·5 (0·3)
	$\mathrm{EP}_{q=1}$	-0·2 (0·2)	0·6 (0·3)	-0·6 (0·3)	0·0 (0·8)	-0·2 (0·3)	0·5 (0·5)	-1·1 (0·3)	-0·2 (0·1)
	$\mathrm{EP}_{q=0·2}$	-1·3 (0·2)	-1·8 (0·3)	-0·8 (0·3)	-1·4 (1·2)	-0·6 (0·2)	-0·8 (0·7)	-0·3 (0·3)	0·2 (0·2)
	GDP	-1·4 (0·2)	-2·1 (0·4)	-0·8 (0·4)	-1·3 (1·1)	-0·6 (0·2)	-0·6 (0·6)	-1·0 (0·3)	-0·1 (0·1)
	Log	-1·4 (0·2)	-1·9 (0·4)	-0·8 (0·3)	-1·3 (1·1)	-0·6 (0·2)	-0·6 (0·6)	-0·6 (0·3)	0·0 (0·2)

W_G, G-Wishart; WCK, Wong et al. (2003); GDP, generalized double Pareto; EP, exponential power; Log: logarithmic.

priors have very similar performances – ranking among top models in all cases. In summary, the experiment illustrates that these three heavy-tailed priors in the scale mixture of uniform family are generally indeed good for high dimensional covariance matrix estimation.

5 Application to multivariate conditional autoregressive models

5.1 Multivariate conditional autoregressive models based on scale mixture of uniform priors

Multivariate conditional autoregressive models (Banerjee et al., 2004) constitute a diverse set of powerful tools for modeling multivariate spatial random variables at areal unit level. Let $W = (w_{ij})_{p_r \times p_r}$ be the symmetric proximity matrix of p_r areal units, $w_{ij} \in \{0, 1\}$, and w_{ii} are customarily set to 0. Then W defines an undirect graph $G_r = (V_r, E_r)$ where an edge $(i, j) \in E_r$ if and only if $w_{ij} = 1$. Let $w_{i+} = \sum_j w_{ij}$, $E_W = \mathrm{diag}(w_{1+}, \ldots, w_{p_r+})$ and $M = (m_{ij}) = E_W - \rho W$. Let $X = (x_1, \ldots, x_{p_r})^{\mathrm{T}}$ denote a $p_r \times p_c$ random matrix where each x_i is a p_c-dimensional vector corresponding to region i. Following Gelfand & Vounatsou (2003), one popular version of the multivariate conditional autoregressive models sets the joint distribution of X as

$$\mathrm{vec}(X) \sim N\{0, (\Omega_c \otimes \Omega_r)^{-1}\}, \quad \Omega_r \mid \rho = E_W - \rho W, \quad \Omega_c \sim \mathrm{W}(b_c, D_c), \tag{11}$$

where Ω_r is the $p_r \times p_r$ column covariance matrix, Ω_c is the $p_c \times p_c$ row covariance matrix, ρ is the coefficient measuring spatial association and is constrained to be between the reciprocals of the minimum and maximum eigenvalues of W to ensure that Ω_r is nonsingular, and b_c and D_c respectively denote the degree of freedom and location parameters of a Wishart prior distribution for Ω_c. The joint distribution in (11) implies the following conditional distribution:

$$x_i \mid x_{-i}, \rho, \Omega_c \sim \mathrm{N}\Big(\sum_{j \in \mathrm{ne}(i)} \rho\, w_{i+}^{-1}\, x_j, w_{i+}^{-1}\Omega_c\Big),$$

where $\mathrm{ne}(i)$ denotes the neighbor of region i, that is, the set of points satisfying $w_{ij} = 1$. Evidently, the two covariance structures (Ω_r, Ω_c) are crucial in determining the effects of spatial smoothing. For the matrix Ω_c, direct application of shrinkage priors can reduce estimation uncertainties as compared to the conjugate Wishart prior in (11). For Ω_r, one common value of ρ for all x_i may limit the flexibility of the model because it assumes the same spatial association for all regions. The recent work of Dobra et al. (2011) uses the G-Wishart framework to provide alternative models. Specifically, the authors recommend the following extensions for modeling (Ω_r, Ω_c):

$$\Omega_r \mid M \sim \mathrm{W}_{G_r}(b_r, M), \quad M \mid \rho = E_W - \rho W, \quad \Omega_c \sim \mathrm{W}_{G_c}(b_c, D_c), \tag{12}$$

where the row graph G_r is fixed and obtained from the proximity matrix W, and the column graph G_c is unknown. For both models in (11) and (12), a prior for ρ was chosen to give higher probability mass to values close to 1 to encourage sufficient spatial dependence. In particular, Dobra et al. (2011) put equal mass on the following 31 values: $\{0,\ 0{\cdot}05,\ 0{\cdot}1,\ \ldots,\ 0{\cdot}8,\ 0{\cdot}82,\ \ldots,\ 0{\cdot}90,\ 0{\cdot}91,\ \ldots,\ 0{\cdot}99\},$. Notice that Ω_r and Ω_c are not uniquely identified since, for any $c > 0$, $\Omega_c \otimes \Omega_r = (c\,\Omega_c) \otimes (\Omega_r / c)$ (Wang & West, 2009). We address this by fixing $\Omega_{r,11} = 1$.

Using the theory and methods for covariance matrix developed in Section 3, we now extend the multivariate conditional autoregressive models (11) using the scale mixture of uniform distributions. We consider the following two extensions for modeling $\Omega_r \in M^+(G_r)$ and $\Omega_c \in M^+$

$$\Omega_r \mid \rho = E_W - \rho W, \quad p(\Omega_c \mid \tau) \propto \prod_{i \geq j} g_c(\omega_{c,ij}/\tau_c), \tag{13}$$

and

$$p(\Omega_r) \quad \propto \prod_{\{(i,j)\in E_r\}\cup\{i=j\in V_r\}} g_r(|\omega_{r,ij} - m_{ij}|/\tau_r)\, 1_{\{\omega_{r,ij}<0\}}, \quad p(\Omega_c \mid \tau_c) \propto \prod_{i \geq j} g_c(\omega_{c,ij}/\tau_c) \tag{14}$$

The first extension (13) places shrinkage priors on Ω_c while leaving the model for Ω_r unchanged. The second extension (14) further shrinks Ω_r towards the matrix $M = E_W - \rho W$ while allowing adaptive spatial smoothing by not constraining Ω_c to be controlled by a common parameter ρ.

One practical advantage of the our model (14) over the model (12) of Dobra et al. (2011) is its flexibility in incorporating prior knowledge. For example, the similarity of spatial neighbors implies that the off-diagonal elements of Ω_r should be constrained to be negative (Banerjee et al., 2004). This point is not addressed by Dobra et al. (2011) as their method is only applicable when the free elements of Ω_r are not truncated. In the scale mixture of uniform framework, this important constraint is easily achieved by truncating each free off-diagonal element in Ω_r to be negative when sampling Ω_r. The functional form of $g_r(\cdot)$ and the shrinkage parameter τ_r can be pre-specified through careful prior elicitation as follows. Using the Gibbs sampler in Section 3.1, we are able to simulate from the prior distribution of Ω_r for fixed $g_r(\cdot)$ and τ_r. These prior draws allow us to choose $g_r(\cdot)$ and τ_r to represent plausible ranges of spatial associations. To specify these ranges, one guideline can be based on the model (11) for which Gelfand & Vounatsou (2003) recommended a prior for ρ that favors the upper range of $\rho \in (0, 1)$. In light of this recommendation, we prefer those g_r and τ_r that increasingly favor values of $\omega_{c,ij}$ close to 1 for any $(i, j) \in E_r$ and $\omega_{c,ii}$ close to w_{i+} for $i \in V_r$. Such choices of priors integrate prior information about spatial associations and allow for varying spatial smoothing parameters across different regions.

5.2 US cancer data

Using our model, we analyze the same real data example studied by Dobra et al. (2011) concerning the application of multivariate spatial models for studying the US cancer mortality rates. The data we analyzed consists of mortality counts for 10 types of tumors recorded for the 48 mainland states plus the District of Columbia for the year 2000. The data were collected by the National Center for Health Statistics. Morality counts below 25 were treated as missing because they are regarded as unreliable records in cancer surveillance community. Let Y_{ij} be the number of deaths in state $i = 1, \ldots, p_r = 49$ for tumor type $j = 1, \ldots, p_c = 10$. Following Dobra et al. (2011), we considered Poisson multivariate loglinear models with spatial random effects:

$$Y_{ij} \mid \eta_{ij} \sim \mathrm{Poi}(\eta_{ij}), \quad \log(\eta_{ij}) = \log(q_i) + \mu_j + X_{ij},$$

where q_i is the population of state i, μ_j is the intercept of tumor type j and X_{ij} is the zero-mean spatial random effect associated with state i and tumor j and has the joint distribution $\mathrm{vec}(X) \sim \mathrm{N}\{0, (\Omega_c \otimes \Omega_r)^{-1}\}$.

We compared the out-of-sample predictive performance of model (13) and (14) against the model (11) of Gelfand & Vounatsou (2003) and model (12) of Dobra et al. (2011). For (11) and (12), we used the same hyper-parameter settings as in Dobra et al. (2011). For (13), we set $g_c(\cdot)$ to be the logarithmic density in (4) and placed standard half-cauchy prior on τ_c in order to expect robust performance for shrinkage estimation of Ω_c as was suggested by the simulation study in Section 4. For (14), we let $g_r(\omega_{r,ij}) \propto \exp\{-|\omega_{r,ij} - m_{ij}|/\tau_r\}1_{\{\omega_{r,ij}<0\}}$ for $i = j$ or $(i, j) \in E_r$ so that Ω_r is centered around $M = W - E_W$ and the similarity of spatial neighbors is ensured. We did not choose heavy-tailed distributions for $g_r(\cdot)$ because the sample size $p_c = 10$ is relatively small for the dimension $p_r = 49$ and a heavy-tailed prior can lead to a posterior distribution of $\omega_{r,ij}$ to be unrealistically small and $\omega_{r,ii}$ to be unrealistically large. We considered $\tau_r \in \{0{\cdot}1,\ 1,\ 10\}$ to assess the prior sensitivity. Finally, we modeled $g_c(\cdot)$ as in model (13).

In order to assess the out-of-sample predictive performance, we replicated the 10-fold cross-validation experiment of Dobra et al. (2011). Specifically, we divided the nonmissing counts of Y into 10 bins. For each bin i, we used the samples from the other 9 bins as observed data and imputed the samples from bin i as missing. To compare different models, we then computed the predictive mean squared error and mean variance as follows

$$\mathrm{MSE} = \frac{1}{|\{(i,j) : Y_{ij} \geq 25\}|} \sum_{\{(i,j):Y_{ij}\geq 25\}} (E(Y_{ij}) - Y_{ij})^2,$$

and

$$\mathrm{VAR} = \frac{1}{|\{(i,j) : Y_{ij} \geq 25\}|} \sum_{\{(i,j):Y_{ij}\geq 25\}} Var(Y_{ij}),$$

where $E(Y_{ij})$ and $Var(Y_{ij})$ are estimated using the posterior sample mean and variance based on the output of the analysis of one of the 10 cross-validation datasets in which Y_{ij} are treated as missing. All results were obtained using a Monte Carlo sample of size 80000 after an initial, discarded burn-in of 80000 iterations.

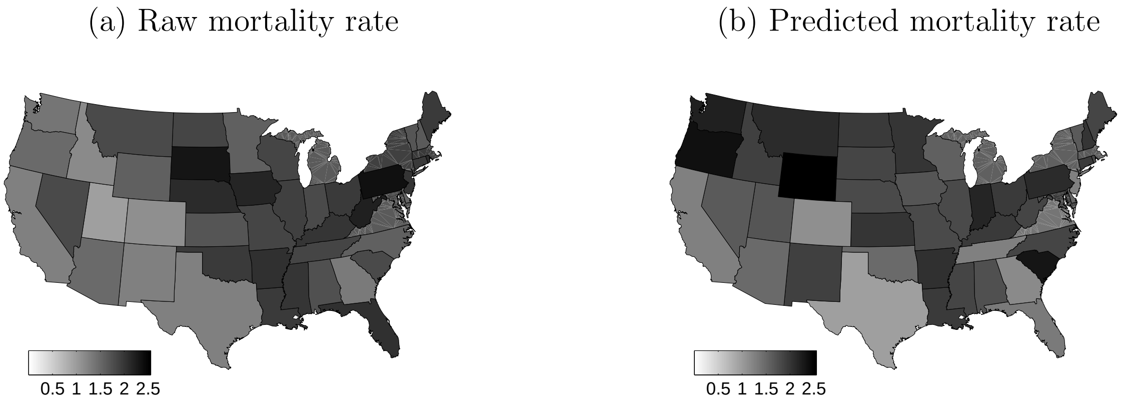
Figure 1 shows the raw and predicted morality rate of colon cancer. Table 3 reports the predictive performance as measured by the mean squared error and mean variance. All methods with shrinkage priors on Ω_c improve the prediction over the standard method using the Wishart prior. Among the shrinkage methods, the logarithmic prior outperforms the G-Wishart prior. Allowing Ω_r to be adaptive by setting $\tau_r = 1$ and 10 can further reduce the mean squared error while maintaining the same predictive variance with the common ρ model. Overall, our results suggest that the models (13) and (14) provide more accurate prediction and narrower credible intervals than the competing methods for this dataset.

To further study the prior sensitivity to the choice of τ_r, we plotted the marginal prior and posterior densities for the free off-diagonal element in Ω_r using samples from the analysis of the first cross-validation dataset. Figure 2 displays the inference for one element under $\tau_r \in \{0{\cdot}1,\ 1,\ 10\}$. Clearly, the marginal posterior distribution depends on the choice of τ_r. This is not surprising because the sample size is small compared to the dimension of Ω_r. The case $\tau_r = 1$ and 10 seems to perform well in this example because the marginal posterior distribution is influenced by the data. The case $\tau_r =$0·1 appears to be too tight and thus is not largely influenced by the data.

On the computing time, the Matlab implementation of model (14) took about 4 hours to complete the analysis of one of the ten cross-validation datasets, while model (12) of Dobra et al. (2011) took about 4 days. Additionally, Dobra et al. (2011) reported a runtime of about 22 hours on a dual-core 2·8 Ghz computer under C++ implementation for a similar dataset of size $p_r = 49$ and $p_c = 11$. As mentioned above, our models based on the scale mixture of uniforms are not only more flexible but also more computationally efficient.

6 Shrinkage prior for linear regression models

In this section we briefly investigate the properties of the shrinkage prior constructed from scale mixture of uniforms for the linear regression models. Recently, shrinkage estimation for linear models have received a lot of attention (Park & Casella, 2008; Griffin & Brown, 2010; Armagan et al., 2011) all of which proceed via the scale mixture



(a) Raw mortality rate

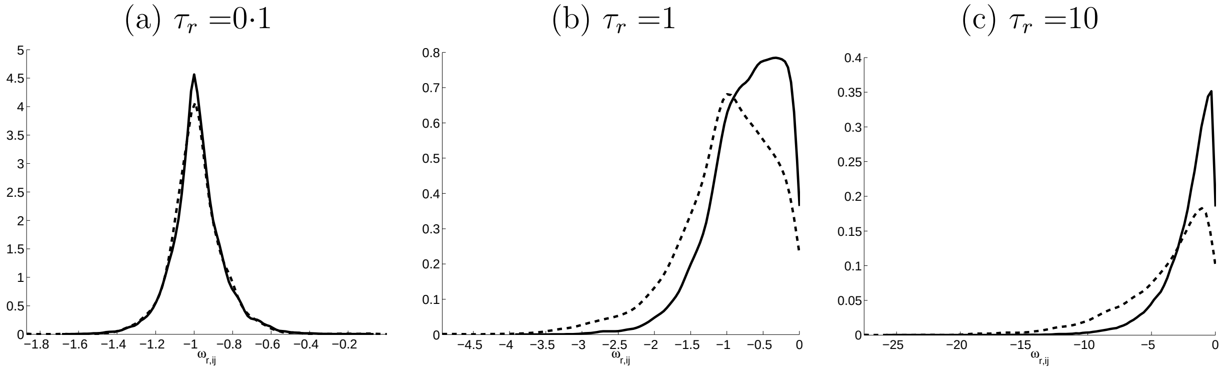
(b) Predicted mortality rate

Figure 1: US cancer mortality map of colon cancer (per 10000 habitants). (a) The raw mortality rate, (b) The predicted mortality rate under model TDE+Log with $\tau_r = 1$.

Table 3: Predictive mean squared error and variance in 10-fold cross-validation predictive performance in the cancer mortality example.

	GV	DLR	Common ρ+Log		TDE+Log	
				τ_r=10	τ_r=1	τ_r=0·1
MSE	3126	2728	2340	2238	2187	2359
VAR	9177	6493	3814	3850	3810	3694

GV: the non-shrinkage model (11) of Gelfand & Vounatsou (2003); DLR: model (12) of Dobra et al. (2011); Common ρ+Log: model (13) under common ρ for Ω_r and logarithmic prior for Ω_c; TDE+Log: model (14) under truncated double-exponential prior for Ω_r with fixed but different τ_r and logarithmic prior for Ω_c.



(a) τ_r =0·1

(b) τ_r =1

(c) τ_r =10

Figure 2: Marginal prior (dashed lines) and posterior (solid lines) densities of one free off-diagonal element in Ω_r from the analysis under model (14) with three different values of τ_r: (a) τ_r =0·1, (b) τ_r =1, (c) τ_r =10.

of normals. Walker et al. (1997) and Qin et al. (1998) were among the first to use the scale mixture of uniform priors for regression models. However, they used this family only for modeling the measurement errors and deriving the corresponding Gibbs sampler. To the best of our knowledge, we are the first to investigate the scale mixture of uniforms as a class of shrinkage priors for regression coefficients. When this paper was nearing completion we were notified of a similar approach in the very recent work Polson & Scott (2011) in which the authors independently propose a similar construction based on mixtures of Bartlett-Fejer kernels for the bridge regression and proceed via a result similar to Theorem 1.

Consider the following version of a regularized Bayesian linear model where the goal is to sample from the posterior distribution

$$p(\beta \mid \sigma, \tau, Y) \propto \exp\{-\frac{1}{2\sigma^2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta)\} \prod_{j=1}^{p} g(\frac{\beta_j}{\sigma\tau})$$

where $g(\cdot)$ is the shrinkage prior and τ is the global shrinkage parameter. Theorem 1 suggests we can introduce latent variable $t = \{t_1, \dots, t_p\}$ such that the joint posterior of (β, t) is given by:

$$p(\beta, t \mid \sigma, \tau, Y) \propto \exp\{-\frac{1}{2\sigma^2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta)\} \prod_{j=1}^{p} \{-g'(t_j)\, 1_{\{\sigma\tau t > |\beta_j|\}}\}$$

The Gibbs samplers are then implemented by (a) simulating β_j from a truncated normal for each j, and (b) block simulating $\{t_1, \dots, t_p\}$ from using the conditional cumulative distribution function in Theorem 2.

We compare the posterior mean estimators under the exponential power prior with q =0·2 and the logarithmic prior to the posterior means corresponding to several other existing priors. These two shrinkage priors are interesting because the exponential power prior is the Bayesian analog of the bridge regression (Park & Casella, 2008) and is challenging for fully posterior analysis using the scale mixture of normals and relatively unexplored before, and the logarithmic prior is a new prior that resembles the class of horseshoe priors that are shown to have some advantages over many existing approaches (Carvalho et al., 2010).

We use the setting of simulation experiments considered in Armagan et al. (2011). Specifically, we generate $n = 50$ observations from $y = x^{\mathrm{T}}\beta + \epsilon, \epsilon \sim \mathrm{N}(0, 3^2)$, where β has one of the following five configurations: (i) β =(1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$^{\mathrm{T}}$, (ii) β(3,3,3,3,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$^{\mathrm{T}}$, (iii) β =(1,1,1,1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0)$^{\mathrm{T}}$, (iv) β =(3,3,3,3,3,0,0,0,0,0,3,3,3,3,3,0,0,0,0,0)$^{\mathrm{T}}$, (v) β =(0·85, ..., 0·85)$^{\mathrm{T}}$, and $x = (x_1, \dots, x_p)^{\mathrm{T}}$ has one of the following two scenarios: (a) x_j are independently and identically distributed standard normals, (b) x is a multivariate normal with $E(x) = 0$ and $\mathrm{cov}(x_j, x_{j'}) = 0{\cdot}5^{|j-j'|}$. The variance is assumed to have the Jeffrey's prior $p(\sigma^2) \propto 1/\sigma^2$. The global shrinkage parameter is assumed to have the conjugate $\tau^{-q} \sim \mathrm{Ga}(1, 1)$ for the exponential power prior with q =0·2, and $\tau \sim \mathrm{C}^{+}(0, 1)$ for the logarithmic prior.

Table 4: Summary of model errors for the simulation study in the regression analysis of Section 6. Median model errors are reported; bootstrap standard errors are in parentheses.

	(a) x_j independent					(b) x_j correlated				
	(i)	(ii)	(iii)	(iv)	(v)	(i)	(ii)	(iii)	(iv)	(v)
GDP^a	2·7 (0·1)	2·2 (0·2)	4·0 (0·2)	3·8 (0·2)	5·7 (0·3)	2·1 (0·1)	2·1 (0·1)	3·2 (0·1)	4·2 (0·3)	4·4 (0·1)
GDP^b	2·8 (0·2)	2·1 (0·2)	4·6 (0·2)	3·8 (0·2)	7·0 (0·2)	1·9 (0·1)	2·0 (0·1)	3·3 (0·2)	4·2 (0·2)	4·7 (0·1)
GDP^c	2·6 (0·1)	2·4 (0·2)	4·4 (0·2)	4·0 (0·2)	6·5 (0·2)	1·9 (0·1)	2·2 (0·1)	3·1 (0·2)	4·3 (0·2)	4·3 (0·1)
HS	2·7 (0·1)	2·1 (0·2)	4·8 (0·2)	3·8 (0·2)	7·3 (0·2)	2·0 (0·1)	2·0 (0·1)	3·3 (0·2)	4·3 (0·2)	4·6 (0·1)
EP$_{q=1}$	3·2 (0·1)	4·0 (0·3)	5·1 (0·3)	4·9 (0·3)	7·3 (0·5)	2·1 (0·1)	2·8 (0·2)	2·8 (0·1)	4·2 (0·3)	3·5 (0·2)
EP$_{q=0\cdot2}$	2·5 (0·1)	2·0 (0·1)	4·7 (0·1)	3·9 (0·3)	7·3 (0·3)	2·0 (0·1)	2·1 (0·1)	3·2 (0·1)	3·9 (0·1)	5·4 (0·2)
Log	2·5 (0·1)	2·5 (0·2)	4·5 (0·2)	4·5 (0·2)	6·4 (0·4)	2·0 (0·1)	2·4 (0·1)	3·0 (0·1)	4·3 (0·1)	4·6 (0·2)

GDP^{a,b,c}, three recommended Generalized double Pareto priors in Armagan et al. (2011); HS, horseshoe; EP, exponential power; Log, logarithmic.

Model error is calculated using the Mahalanobis distance $(\hat{\beta}-\beta)^{\mathrm{T}}\Sigma_X(\hat{\beta}-\beta)$ where Σ_X is the covariance matrix used to generate X.

Table 4 reports the median model errors and the bootstrap standard errors based on 100 datasets for each case. Results for cases other than the exponential power prior with q =0·2 and the logarithmic prior are based on the reported values of Armagan et al. (2011). Except for model (iii) and (v) in the correlated predictor scenario, the exponential power prior with $q = 1$ is outperformed by other methods. The performances of the exponential power prior with q =0·2 and the logarithmic prior are comparable with those of the generalized Pareto and the horseshoe priors.

7 Conclusion

The scale mixture of uniform prior provides a unified framework for shrinkage estimation of covariance matrices for a wide class of prior distributions. Further research on the scale mixture of uniform distributions is of interest in developing theoretical insights as well as computational advances in shrinkage prior estimation for Bayesian analysis of covariance matrices and other related models. One obvious next step is to investigate the covariance selection models that encourage exact zeros on a subset of elements of Ω under the scale mixture uniform priors. Such extensions can potentially combine the flexibility of the scale mixture of uniform priors and the interpretation of the graphs implied by exact zero elements. Another interesting research direction is the generalization of the basic random sampling models to dynamic settings that allow the covariance structure to be time-varying. Such models are useful for analyzing high-dimensional time series data encountered in areas such as finance and environmental sciences. We are current investigating these extensions and we expect the Gibbs sampler developed in Section 3.1 to play a key role in model fitting in these settings.

Acknowledgements

The authors thank Abel Rodriguez and James G. Scott for very useful conversations and references. NSP gratefully acknowledges the NSF grant DMS 1107070.

Appendix

Details of sampling algorithm in Section 3.1

The joint distribution of (l_{12}, d_1, d_2) is:

$$p(d_1, d_2, l_{21} \mid -) \propto d_1^{n/2+1} d_2^{n/2} \exp[-\frac{1}{2}\mathrm{tr}\{s_{11}d_1 + s_{22}(l_{21}^2 d_1 + d_2) + 2s_{21}d_1 l_{21}\}]\, 1_{\{\Omega_{e,e}\in\mathcal{T}\}}.$$

Clearly, the full conditional distribution for d_1, d_2 and l_{21} are given by

$$d_1 \sim \textsc{Ga}\{n/2+2, (s_{11} + s_{22}l_{21}^2 + 2s_{21}l_{21})/2\}\, 1_{\{\Omega_{e,e}\in\mathcal{T}\}}\ ,$$

$d_2 \sim \textsc{Ga}(n/2+1, s_{22}/2)\, 1_{\{\Omega_{e,e}\in\mathcal{T}\}}$ and $l_{21} \sim \mathrm{N}\{s_{21}/s_{22}, 1/(s_{22}d_1)\}\, 1_{\{\Omega_{e,e}\in\mathcal{T}\}}$, respectively. To identify the truncated region $\mathcal{T}$, recall

$$\Omega_{e,e} = A + B, \quad A = \begin{pmatrix} d_1 & d_1 l_{21} \\ d_1 l_{21} & d_1 l_{21}^2 + d_2 \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

The set $\mathcal{T} = \{|\omega_{ij}| < t_{ij}\} \cap \{|\omega_{ii}| < t_{ii}\} \cap \{|\omega_{jj}| < t_{jj}\}$ can be written as

$$\{|d_1 + b_{11}| < t_{ii}\} \cap \{|d_1 l_{21} + b_{21}| < t_{ij}\} \cap \{|d_1 l_{21}^2 + d_2 + b_{22}| < t_{jj}\}. \qquad (15)$$

Given $\{B, t_{ii}, t_{ij}, t_{jj}\}$, (15) gives straightforward expressions for the truncated region of each variable in (d_1, d_2, l_{21}) conditional on the other two.

Sampling a univariate truncated normal can be carried out efficiently using the method of Robert (1995), while sampling a truncated gamma is based on the inverse cumulative distribution function method.

References

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, pp. 99–102.

Armagan, A. (2009). Variational bridge regression. *Proceedings of the 12th International Confe- rence on Artificial Intelligence and Statistics (AISTATS)* **5**.

Armagan, A., Dunson, D. & Lee, J. (2011). Generalized double Pareto shrinkage. *ArXiv e-prints* .

Atay-Kayis, A. & Massam, H. (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika* **92**, 317–35.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004). *Hierarchical Modeling and analysis of Spatial data*. Boca Raton: Chapman & Hall.

Barnard, J., McCulloch, R. & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer Series in Statistics, New York: Springer, 2nd ed.

Berger, J. O. & Berliner, L. M. (1986). Robust bayes and empirical bayes analysis with ϵ-contaminated priors. *The Annals of Statistics* **14**, 461–486.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Carvalho, C. M. & Scott, J. G. (2009). Objective bayesian model selection in gaussian graphical models. *Biometrika* **96**, 497–512.

Daniels, M. J. & Kass, R. E. (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* **94**, pp. 1254–1263.

Daniels, M. J. & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173–1184.

Dobra, A., Lenkoski, A. & Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association (to appear)* .

Fan, J., Feng, Y. & Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *Annals of Applied Statistics* **3**, 521–541.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, vol. II. New York: John Wiley & Sons, 2nd ed.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Gelfand, A. E. & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11–15.

Griffin, J. E. & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.

Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. & West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

Liechty, J. C., Liechty, M. W. & Müler, P. (2004). Bayesian correlation estimation. *Biometrika* **91**, 1–14.

Liechty, M. W., Liechty, J. C. & Müller, P. (2009). The Shadow Prior. *Journal of Computational and Graphical Statistics* **18**, 368–383.

Mitsakakis, N., Massam, H. & Escobar, M. (2010). A Metropolis-Hastings based method for sampling from G-Wishart distribution in Gaussian graphical models. Tech. rep., University of Toronto.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

Polson, N. G. & Scott, J. G. (2011). The Bayesian Bridge. *ArXiv e-prints* .

Qin, Z., Walker, S. & Damien, P. (1998). Uniform scale mixture models with application to Bayesian inference. Working papers series, University of Michigan Ross School of Business.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125. 10.1007/BF00143942.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

Roverato, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**, 391–411.

Walker, S., Damien, P. & Meyer, M. (1997). On scale mixtures of uniform distributions and the latent weighted least squares method. Working papers series, University of Michigan Ross School of Business.

Wang, H. (2011). The bayesian graphical lasso and efficient posterior computation. Working papers series, University of South Carolina.

Wang, H. & Carvalho, C. M. (2010). Simulation of hyper-inverse wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics* **4**, 1470–1475.

Wang, H. & West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96**, 821–834.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, pp. 646–648.

WONG, F., CARTER, C. & KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–30.

YANG, R. & BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics* **22**, pp. 1195–1211.

YUAN, M. & LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.